基于密度峰值优化的谱聚类算法 *

薛丽霞，孙　伟，汪荣贵，杨　娟[†]，胡　敏

(合肥工业大学 计算机与信息学院，合肥 230009)

摘　要：针对经典谱聚类算法无法自适应确定聚类数目、以及在处理大数据量的聚类问题时效率不高的问题，提出了一种基于密度峰值优化的谱聚类算法。该方法首先计算数据对象的局部密度，以及每个数据对象与较其他数据对象的最小距离，并依据一定的规则自适应产生初始聚类中心，确定聚类数目；其次，使用 Nyström 抽样来降低特征分解的计算复杂度以达到提高谱聚类算法的效率。实验结果表明，该方法能够准确地得到聚类数目，并且有效提高了聚类的准确率和效率。

关键词：谱聚类；密度峰值；密度聚类；自适应；Nyström 抽样

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.01.0019

Spectral clustering based on density peak value optimization

Xue Lixia, Sun Wei, Wang Ronggui, Yang Juan[†], Hu Min

(*School of Computer & Information*, *Hefei University of Technology*, *Hefei* 230009, *China*)

Abstract: To deal with the problem that classical spectral clustering algorithms are unable to determine the number of clusters automatically, and low efficiency in processing large amount of data with. This paper proposes a spectral clustering algorithm based on the optimization of density peak value. The method firstly calculates the local density of data object and the minimum distance between each data object and other data objects. Adaptive clustering algorithm is generated to determine the number of clusters and to optimize the number of clusters according to certain rules. Secondly, adopting Nyström sampling can reduce the time complexity of characteristic decomposition and improve the efficiency of the algorithm. The experimental results show that this method can accurately obtain the number of clusters and effectively improve the accuracy and efficiency of clustering effectively.

Key words: spectral clustering; density peak; density clustering; adaptive; Nyström sampling

0　引言

聚类，顾名思义，就是将数据分类到不同的类或者簇，使得同一个簇中的数据对象具有较高的相似性，而属于不同的簇的数据对象之间存在较大的相异性。目前，许多聚类算法已经被研究人员提出，比如 K-means 算法、EM 算法、FCM 算法等。这些传统的聚类算法对凸形样本空间聚类效果较好，但对于非凸形结构的数据集，聚类效果则不是很好。近年来，谱聚类算法[1-3]逐渐发展成较为重要的聚类算法之一，该算法是一种基于谱图理论的新型聚类分析方法。在复杂的样本空间的聚类中，谱聚类算法通常会表现出更好的聚类性能。

谱聚类算法通过计算相似度矩阵（或亲和矩阵）以及拉普拉斯矩阵并对拉普拉斯矩阵进行特征分解，利用特征向量将原始数据点映射到一个低维的特征空间，在该特征空间中，数据的分布结构更加明显，可以使用经典的聚类算法如 K-means 进行聚类。由于使用 K-means 对特征空间聚类，谱聚类算法需要事先指定聚类中心数目，并且在聚类过程中通常随机选取初始聚类中心，初始聚类中心以及聚类中心数目的选取会影响谱聚类算法的聚类结果。另外，谱聚类算法在对拉普拉斯矩阵进行特征分解过程中，计算复杂度较高，不适用处理大规模数据。

文献[4]提出一种非线性降维算法，该算法能够自动确定聚类的数目，但计算复杂度较高，且得到的结果差强人意。文献[5]提出一种基于本征间隙确定聚类数目的谱聚类方法，该方法利用本征间隙刻画向量特征值之间差，通过第一个极大本征间隙出现的位置来自动确定类个数，但该算法对于不均匀的数据分布和高维数据，容易出现类估计错误和分类准确率低的问题。

收稿日期：2018-01-09；**修回日期：**2018-03-05　　**基金项目：**国家自然科学基金资助项目（61672202）

作者简介：薛丽霞（1976-），女，四川西昌人，副教授、硕导，博士，主要研究方向为数字图像处理和数据挖掘（51003239@qq.com）；孙伟（1993-），男，硕士研究生，主要研究方向为数字图像处理和机器学习；汪荣贵（1966-），男，安徽池州人，教授，博导，博士，主要研究方向为数字图像处理、人工智能和数据挖掘；杨娟（1983-），女（通信作者），辽宁沈阳人，讲师，博士，主要研究方向为数字图像处理、人工智能；胡敏（1967-），女，安徽淮北人，教授，硕导，博士，主要研究方向为数字图像处理和数据挖掘.

文献[6]提出一种基于人工免疫确定聚类数目的谱聚类算法，该算法通过模拟抗体的克隆选择机制和免疫系统的应答系统，使得聚类数目可以自适应调整，但算法中的阈值参数需要多次实验确定，阈值不同导致实验结果有较大差异。文献[7]提出一种基于自然邻的自适应谱聚类算法。该算法利用自然邻产生的局部密度信息和近邻关系对高斯函数进行修正，并自适应获取相应的聚类数目，解决了人工指定聚类数目的缺陷，该算法聚类效果较好，但计算复杂度较高，不适用与大规模数据。

本文针对谱聚类算法存在的潜在缺陷，提出利用密度峰值聚类算法[8,9]改进的谱聚类算法——DP-SC（density peak optimized spectral clustering）。DP-SC 算法通过找出数据集中局部密度较大并且与高密度点之间的距离较大的数据点作为聚类的初始聚类中心，由此得到聚类数目，这样解决了谱聚类算法需要人工指定聚类数目以及随机初始化聚类中心问题；另外，在对拉普拉斯矩阵进行特征分解过程中，引入 Nyström 逼近方法[10]，来降低谱聚类算法的计算复杂度。

1　谱聚类算法

谱聚类算法是一种建立者图论中图谱理论基础上的新型聚类算法，其本质是利用谱松弛方法将聚类问题转换为图的最优划分问题。假定待聚类数据集 $X=\{x_1,x_2,\cdots,x_N\}\epsilon R^d$ 中的每一个样本点看做无向图中的顶点，记为 V，根据样本点间的相似度将顶点之间的边 E 赋权值得到相似度矩阵 W，由此构造了一个基于样本间相似度的无向加权图 $G=(V,E,W)$。谱聚类算法可以归纳为以下四个基本步骤：

a)根据待聚类数据集 X，生成图的相似度矩阵 W，其中每个元素 w_{ij} 可以用高斯核函数来表示，即

$$w_{ij}=\exp\left(-\frac{d\left(x_i-x_j\right)}{2\sigma^2}\right)\tag{1}$$

其中: $d\left(x_i-x_j\right)$ 表示数据点 x_i 和 x_j 之间的距离；σ 为尺度参数。尺度参数 σ 起着极为重要的作用。不同的尺度参数的选取可能会导致不同的聚类结果。本文算法采用文献[11]提出的高斯核函数法。

b)计算 Laplacian 矩阵 $L=D^{-1/2}WD^{-1/2}$，其中 D 为度矩阵，每一个元素 d_{ij} 满足

$$d_{ij}=\sum_j w_{ij}\tag{2}$$

c)对 L 进行特征分解，得到前 k 个特征向量，并构建特征向量空间；

d)利用经典聚类方法如 K-means 对特征向量空间中的特征向量进行聚类。

在上述步骤中，谱聚类算法通常必须给出指定的聚类中心数目 k，这一点往往很难精确给出，因为待分类数据集中的数据往往是无序的，聚类中心数目很难确定。当 k 比较大时，选取的 k 个特征向量不一定都包含聚类信息，从而导致聚类结果出现偏差。

2　密度峰值聚类算法

CFSFDP 算法[8]是 Rodriguez 和 Laio 于 2014 年在 Science 杂志提出的一种基于密度的新型聚类算法。密度峰值聚类算法的核心思想在于:聚类中心被具有较低局部密度的邻居点包围，并且与具有更高密度的任何点之间有相对较大的距离。在文献[8]中，对于待聚类数据集 $X=\left\{x_i\,|\,x_i\epsilon R^d,i=1,2,\cdots,N\right\}$ 中的每一个数据点 x_i，密度峰值聚类算法都需要计算两个关键的参数，即局部密度 ρ_i 和与高局部密度点之间的距离 δ_i。

a）局部密度 ρ_i 的定义如下：

$$\rho_i=\sum_j\chi\left(d_{ij}-d_c\right)\tag{3}$$

其中: $\chi(x)=\begin{cases}1, & x<0\\ 0, & x\ge 0\end{cases}$

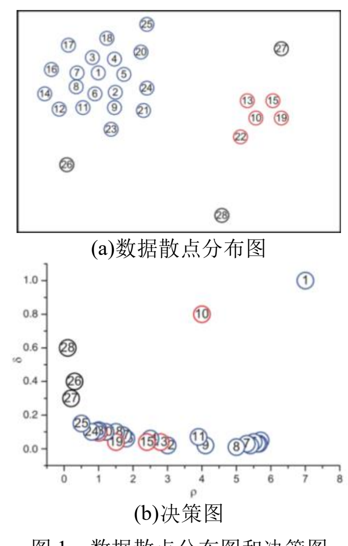
参数 d_c 为截断距离，d_c 取值通常都需要人工指定，本文算法采用文献[12]的做法自适应的获取 d_c。公式 3）目的是为了找到待聚类数据集 X 中与数据点 x_i 之间距离小于 d_c 的数据点个数。

b）与高密度点之间的距离 δ_i 定义如下：

$$\delta_i=\min_{j:\rho_j>\rho_i} d_{ij}\tag{4}$$

其中，对于密度最大的数据点，可以得到 $\delta_i=\max_j d_{ij}$。

根据式（3）（4），对于待聚类数据集 X 中的每一个数据点 x_i，可计算二元对 (ρ_i,δ_i)，并构造以 ρ 为横坐标、δ 为纵坐标的决策图（decision graph）。如图 1 所示，共包括 28 个二维数据点。易知，第 1 号和第 10 号数据点都具有较大的 ρ 和 δ，因而可选择这两个数据点作为聚类中心。另外，对于编号为 26，27，28 的三个数据点在数据集 S 中为异常点，他们都具有共同的特点：δ 值较大，但 ρ 值较小。



(a)数据散点分布图

(b)决策图

图 1　数据散点分布图和决策图

3　基于密度峰值优化的谱聚类算法

针对谱聚类算法中存在随机选择初始聚类中心以及事先指定聚类中心数目的问题，本文结合密度峰值聚类算法提出了一种改进的谱聚类算法——DP-SC（density peaks optimized spectral clustering）聚类算法。DP-SC 算法的主要思想为：初始聚类中心数目不再人工选取，而是基于密度峰值对聚类中心的附近的数据密集程度及与其他聚类中心的距离进行衡量，自适应获得初始聚类中心与聚类数目，使得算法的鲁棒性更强。

3.1　初始聚类中心

通常对于聚类中心的选择，应当遵循如下原则，即：应该尽量让聚类中心反映整体数据集的密集程度，避免聚类中心选取到异常点。DP-SC 算法利用 CFSFDP 算法提出的局部密度 ρ 与距离 δ，并进一步优化可有效确定聚类中心以及聚类中心数目。

为了能够更好地选择初始聚类中心并确定聚类数目，本文采取文献[9]中提出的 Fuzzy-CFSFDP 算法的优化方法。Fuzzy-CFSFDP 算法对 CFSFDP 算法的优化是基于如下公式：

$$EC_i = (\delta_i) \geq 2\sigma(\delta_i) \tag{5}$$

其中，EC_i 代表期望的聚类中心；$\sigma(\delta_i)$ 是根据式（4）计算得到的所有距离的标准差。根据 CFSFDP 算法，聚类中心与其他聚类中心之间有着较大的距离，因此数据集中的其他点的距离将小于 $2\sigma(\delta_i)$。但对于异常点，由于其具有较大的 δ 值而局部密度 ρ 较小，仅通过上式很难将异常点从期望的聚类中心中分离。

为了能够将异常点准确地分离，Fuzzy-CFSFDP 算法使用如下公式：

$$LC_i = EC_i \geq \mu(\rho_i) \tag{6}$$

其中：LC_i 是去除异常点后的局部聚类中心；$\mu(\rho_i)$ 是局部密度 ρ_i 的均值。

通过结合式（5）（6），本文得到的局部聚类中心具有如下特点：比邻居点具有更高的局部密度 ρ 以及更大的 δ 值。之后需要对局部聚类中心进行合并。如果各个局部聚类中心之间最小距离如果小于截断距离 d_c，那么将其合并成一个聚类中心。最终局部聚类中心合并完成后可得到全局聚类中心。更进一步，全局聚类中心数目即为算法所需的聚类数目。

3.2　算法性能优化

谱聚类算法通常只适用于规模较小的数据集，因为在其聚类过程中，存储相似度矩阵需要的空间复杂度为 $O(n^2)$；对拉普拉斯矩阵进行特征分解时，需要的时间复杂度一般为 $O(n^3)$。为了降低谱聚类算法的计算复杂度，本文采用 Nyström 逼近方法[10]优化谱聚类算法。Nyström 方法是由 Delves 和 Mohamed 于 1985 年提出的一种用来近似逼近数值积分中的积分算子的数字逼近技术。该方法实质是用小样本来近似逼近整个数据集。将待聚类数据集 X 划分为两部分，一部分为随机抽样得到的 m $(m \ll N)$ 个样本点，另一部分为剩余的 N-m 个数据点。式 7）给出了 Nyström 逼近方法的矩阵表示：

$$W = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \tag{7}$$

其中：$A \in R^{m \times m}$ 为抽样点间的相似度矩阵，且 $A = U\Lambda U^T$；$B \epsilon R^{(N-m)\times m}$ 为抽样点和剩余点之间的相似度矩阵；$C \in R^{(N-m)\times(N-m)}$ 为剩余点间的相似度矩阵。

令 $\bar{U}$ 表示 W 的近似特征向量，通过 Nyström 扩展可得到：

$$\bar{U} = \begin{bmatrix} U \\ B^T U \Lambda^{-1} \end{bmatrix} \tag{8}$$

相应地，令 $\hat{W}$ 表示近似的 W，则有

$$\hat{W} = \bar{U}\Lambda\bar{U}^T = \begin{bmatrix} U \\ B^T U \Lambda^{-1} \end{bmatrix} \Lambda \begin{bmatrix} U^T & \Lambda^{-1} U^T B \end{bmatrix}$$

$$= \begin{bmatrix} U\Lambda U^T & B \\ B^T & B^T A^{-1} B \end{bmatrix} = \begin{bmatrix} A & B \\ B^T & B^T A^{-1} B \end{bmatrix} \tag{9}$$

Nyström 逼近方法用 $B^T A^{-1} B$ 来间接计算矩阵 C。由于 $n \ll N$，因此利用 $B^T A^{-1} B$ 近似逼近 C，大大降低了问题求解的复杂度。

3.3　算法流程

DP-SC 具体算法如下所示：

输入：待聚类数据集 $X = \{x_1, x_2, \cdots, x_N\}$，采样的数目 m（$m \ll N$）

输出：聚类产生的 k 个类簇

a)利用式（1）计算相似度矩阵 W，并且根据式(3）计算数据点 x_i 的局部密度 ρ_i，以及根据式(4)计算数据点 x_i 的距离 δ_i；

b)根据式(5)(6)得到局部聚类中心后，并对其进行合并，从而获取初始的聚类中心以及聚类数目 k；

c)从数据集 X 中随机选取 m 个抽样点，并依据步骤 1 得到的相似度矩阵 W，计算抽样点间的相似度矩阵 A，抽样点和剩余点之间的相似度矩阵 B；

d)在矩阵 A 和 B 的基础上，利用式(2）计算度矩阵 D，并对 A 和 B 归一化处理（具体的计算公式和原理见文献[2]）；

e)利用归一化后的 A 和 B，计算矩阵 $Q = A + A^{-1/2} B B^T A^{-1/2}$，并对矩阵 Q 对角化，得到总体相似度矩阵的正交特征向量；

f)选取前 k 个特征值对应的特征向量，构建特征向量空间；

g)对特征向量空间的每一行进行规范化，将规范化后的每一行看做待聚类的一个样本点，利用 K-means 聚类算法对该特征向量空间进行聚类。

4　实验分析

为了验证本文提出的 DP-SC 算法的有效性,分别与 NJW 算法[13]、基于 Nyström 抽样的谱聚类算法[14]（简称 NS-SC 算法）以及文献[15]中提出的直接改进距离度量来改变相似度矩阵的方法(简称 SCDL 算法)；其中 NJW、NS-SC 和 SCDL 算法都需要人工指定正确的聚类数目，而本文提出的 DP-SC 算法为自动生成聚类数目。随机抽样个数为待聚类数据集样本数的 10%。实验数据集来源于文献[8]中的 Aggregation 数据集、D31 数据

集和 R15 数据集以及 UCI 机器学习库中的 sonar、dermatology、Wine、Glass、abalone 和 Iris 数据集，他们的具体信息如表 1 所示。本文的实验平台为：Matlab 7.12，操作系统为 Windows 7 64 bit，CPU 为双核 2.60 GHz，RAM 为 4 GB。

表 1　实验数据集描述

数据集	实例数	属性数	类别数
Aggregation	788	2	6
sonar	208	60	2
D31	3100	2	31
dermatology	366	33	6
R15	1500	2	15
Wine	178	13	3
Glass	214	9	6
abalone	4177	8	3
Iris	150	4	3

实验中，采用常用的准确率对算法的聚类结果进行评价[16]。准确率（accuracy）为

$$Acc=\frac{\sum_{i=1}^{N}\varphi(y_i,c_i)}{N} \tag{10}$$

其中：N 为待聚类数据集的数据数目；$\varphi(x,y)$ 为一个函数，当 $x=y$ 时，函数值为 1，否则函数值为 0；y_i 和 c_i 分别表示真实的类别标签和由算法得到的类别标签。显然，当 Acc 值越大时，聚类的效果就越好。

由图 2 和 3 可知，本文提出的 DP-SC 算法在上述 9 种数据集中，准确率更高。根据上面实验结果可以总结出：①本文提出的 DP-SC 算法利用密度峰值算法优化初始聚类中心并自适应确定聚类数目 k，避免了人工指定聚类数目，使得算法在利用 k-means 对特征向量进行聚类时，鲁棒性更强；②由于能够得到准确的聚类数目，使得选取的 k 个特征向量能更好地包含聚类信息；③数据集的聚类数目较大时，DP-SC 算法能够更快的收敛到全局最优解，但相比于 NS-SC 算法，本文算法的计算复杂度还是会高一些，主要是因为本文提出的 DP-SC 算法在自适应确定聚类数目的同时，一定程度上增加了算法的复杂度。

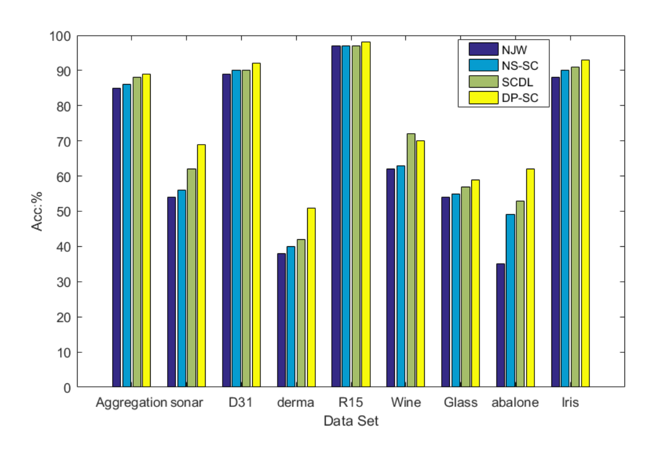


图 2　本文方法与其他算法在不同数据集上聚类准确率对比

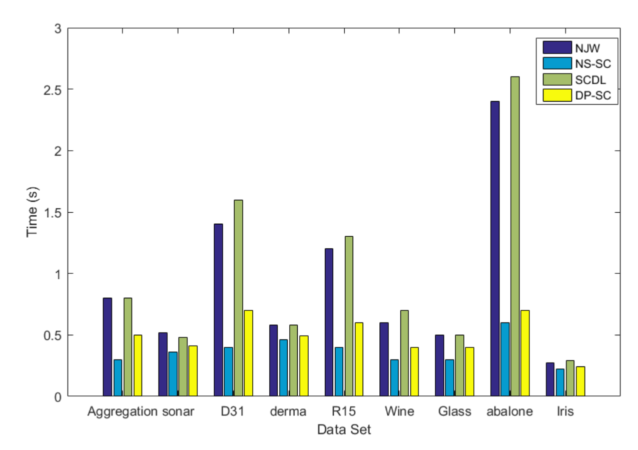


图 3　本文方法与其他算法在不同数据集上运行时间对比

5　结束语

本文基于一种新型密度聚类算法-CFSFDF 算法与当前应用最为广泛的谱聚类算法结合提出一种密度峰值优化的谱聚类算法（DP-SC），能够优化初始聚类中心和自适应确定聚类数目 k，避免人工指定聚类数目；在计算相似度矩阵时，采用基于共享近邻的自适应高斯核函数法，无须事先设定尺度参数 σ；利用 Nyström 逼近方法降低特征向量求解的计算复杂度，该方法在理论上能够提高谱聚类的聚类准确率和效率。下一步的工作是对算法进一步的优化，降低算法的计算复杂度并提高算法的鲁棒性。

参考文献：

[1] Shi J, Malik J. Normalized Cuts and Image Segmentation [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2000, 22 (8): 888-905.

[2] 丁世飞，贾洪杰，史忠植. 基于自适应 Nyström 采样的大数据谱聚类算法 [J]. 软件学报, 2014, 25 (9): 2037-2049.

[3] 王英博，马菁，宋晓倩. 基于最优投影的半监督谱聚类算法 [J]. 计算机应用研究, 2018, 35 (1): 97-100.

[4] Luxburg, Ulrike. A tutorial on spectral clustering [J]. Statistics & Computing, 2007, 17 (4): 395-416.

[5] 孔万增，孙志海，杨灿，等. 基于本征间隙与正交特征向量的自动谱聚类 [J]. 电子学报, 2010, 38 (8): 1880-1885+1891.

[6] 郭凯，李海芳，王会青. 一种人工免疫的自适应谱聚类算法 [J]. 小型微型计算机系统, 2013, 34 (4): 856-859.

[7] 牛海燕，陈笑蓉. 自适应的模糊谱聚类算法在文本聚类中的应用 [J]. 贵州大学学报：自然科学版, 2015, 32 (6): 75-78.

[8] Rodriguez A, Laio A. Machine learning. clustering by fast search and find of density peaks. [J]. Science, 2014, 344 (6191): 1492.

[9] Mehmood R, Bie R, Dawood H, *et al.* Fuzzy clustering by fast search and find of density peaks [C]// Proc of International Conference on Identification, Information, and Knowledge in the Internet of Things. 2016: 785-793.

[10] San J, Mehryar M, Ameet T. Sampling methods for the Nyström method [J]. Journal of Machine Learning Research, 2012, 13: 981-1006

[11] Zhang X, Li J, Yu H. Local density adaptive similarity measurement for spectral clustering [J]. Pattern Recognition Letters, 2011, 32 (2): 352-358.

[12] Lv Zhenghua, Wang Junhua, Shi Xia, *et al.* Clustering by fast searching density peaks based on parameter optimization [C]// Proc of International Conference on Mechatronics, Materials, Chemistry and Computer Engineering. 2017.

[13] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: analysis and an algorithm [C]// Proc of International Conference on Neural Information Processing Systems: Natural and Synthetic. Cambridge: MIT Press, 2001: 849-856.

[14] Fowlkes C, Belongie S, Chung F, Malik J. Spectral grouping using the Nyström method [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2004, 26 (): 214-225.

[15] 牛科，张小琴，贾郭军. 基于距离度量学习的集成谱聚类算法 [J]. 计算机工程, 2015, 41 (1): 207-210.

[16] Chen Wenyen, Song Yangqiu, Bai Hongjie, *et al.* Parallel spectral clustering in distributed systems [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2011, 33 (3): 568-586.